ESTIMATION OF DIRECTION OF ARRIVAL USING MATCHING PURSUIT AND ITS APPLICATION TO SOURCE SEPARATION

Yasuhiro Oikawa[*] *and Yoshio Yamasaki*

Global Information and Telecommunication Institute, Waseda University
1-3-10 Nishi-Waseda, Shinjuku-ku, Tokyo 169-0051, Japan
{oikawa, yamasaki}@giti.waseda.ac.jp

ABSTRACT

In this paper, we describe a new blind source separation (BSS) method that uses spatial information derived from the direction of arrival (DOA) estimates of each direct and reflected sound. The method we proposed has the following steps: (1) each DOA is estimated using matching pursuit and re-optimized after each new DOA is estimated, (2) using these DOA estimates, the mixing matrix is also estimated and the inverse of the mixing matrix is used to separate the mixture signals. Our experiments yielded a better signal separation with the new method than with the conventional frequency domain independent component analysis (ICA) based BSS method.

1. INTRODUCTION

Blind source separation (BSS) has recently been studied by many researchers. Many methods based on independent component analysis (ICA) have been proposed [1], [2], [3], [4], [5], [6]. ICA is used to estimate the unmixing matrix and to separate mixture signals into independent components, assuming that the source signals are independent. Non-ICA based BSS methods have also been proposed [7], [8]. Sparse decomposition with matching pursuit and applying it to BSS was proposed by Gribonval [8] where he computed the sparse decomposition of stereo audio signals with a matching pursuit type algorithm and found the parameters of the atoms of decomposition were clustered. Estimates of sources were then recovered by partial reconstruction using only the appropriate atoms of decomposition. For instantaneous mixtures or for convolved mixtures consisting of short impulse responses, these methods are very effective in separating sources. However, in the case where the mixture is a convolved mixture and impulse responses are long, which is commonly true in the real world, they perform poorly and the separation is not enough [9].

In the real world, the transfer function between the source and microphone has long impulse response with many reflected sounds. If we could estimate all reflected sounds, it would be possible to estimate the complete impulse response and to separate sounds. As it is difficult to estimate all reflected sounds, it would be more effective to consider the spatial information and estimate direct and main reflected sounds to establish a source separation system. In the BSS methods that have been suggested, spatial information has not only been considered for instantaneous mixtures but also for convolved mixtures in the real world.

In this paper, we propose a BSS method that uses spatial information derived from the results of direction of arrival (DOA) estimates for direct and early reflected sounds. We need to find many DOAs to estimate the mixing system. However it is impossible to find true DOAs using conventional beam forming techniques when the number of sources exceeds that of microphones. We suggest a new DOA estimates technique, which is using a matching pursuit algorithm, and it is possible to find true DOAs even if the number of sources exceeds that of microphones. The basic outline of our algorithm is as follows. We first find the normalized power of the array output, $P(\theta)$, as a function of the DOA, θ. Then to estimate the DOA of direct and indirect (reflected) signals we apply a matching pursuit algorithm, which includes a re-optimization of the DOAs at each iteration step. The sounds coming from different DOAs are then classified into a small set of sources. We then form estimates of the impulse responses for each source and microphone combination from these classified DOAs. The separated source signals are obtained by filtering the observations with the inverse of the mixing matrix estimate.

We compared our method with the conventional frequency domain ICA based BSS method [5] using two sources, two microphones, and a convolved mixture. Our experiments yielded better signal separation for the new method than that for conventional frequency domain ICA based BSS.

2. DOA ESTIMATION

The DOA estimation we did consists of the following steps. We first separately calculate the normalized power of the array output, $P(\theta)$, for each frequency bin using the Delay-and-Sum method [10]. We then average $P(\theta)$ over all frequency bins. Finally, we perform peak picking using a matching pursuit algorithm to estimate the DOA over all frequency bands. The matching pursuit algorithm includes, after each iteration step, a re-optimization of all DOAs found thus far. Its main characteristics is that it is possible to find true DOAs when the number of sources exceeds that of microphones. We will discuss these steps in more detail in the following subsections.

2.1. Calculation of power of array output

The power of the Delay-and-Sum array output is calculated as

$$P(\theta) = \mathbf{d}(\theta)^H \mathbf{R} \mathbf{d}(\theta), \quad (1)$$

where $\mathbf{d}(\theta)$ is the steering vector:

$$\mathbf{d}(\theta) = [1, \exp(-j\omega\tau), \cdots, \exp(-j\omega(M-1)\tau)]^T. \quad (2)$$

[*] The author performed the work as a guest researcher at KTH.

Here, M is the number of microphones, $\tau = \frac{d \sin\theta}{c}$, d is the distance between microphones, c is the velocity of sound, and $\mathbf{R}$ is the covariance matrix of the array outputs $\mathbf{x}(t)$ i.e.:

$$\mathbf{R} = E[\mathbf{x}(t)\mathbf{x}(t)^H]. \tag{3}$$

For K sounds (i.e., K different DOAs) and two microphones, the observed signals are

$$\mathbf{X}(\omega,t) = \begin{bmatrix} X_1(\omega,t) \\ X_2(\omega,t) \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^{K} H_{1k}(\omega) S_k(\omega,t) \\ \sum_{k=1}^{K} H_{2k}(\omega) S_k(\omega,t) \end{bmatrix}, \tag{4}$$

where H_{1k} and H_{2k} are the respective transfer functions between the k'th sound and each microphone. The covariance matrix is

$$\mathbf{R}(\omega) = E[\mathbf{X}(\omega,t)\mathbf{X}(\omega,t)^H] = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix}, \tag{5}$$

$$r_{11} = E\big[|\sum_{k=1}^{K} H_{1k} S_k|^2\big], \tag{6}$$

$$r_{12} = E\big[\sum_{k=1}^{K} H_{1k} H_{2k}^* |S_k|^2 + \sum_{l \neq k} H_{1k} H_{2l}^* S_k S_l^*\big], \tag{7}$$

$$r_{21} = r_{12}^*, \tag{8}$$

$$r_{22} = E\big[|\sum_{k=1}^{K} H_{2k} S_k|^2\big], \tag{9}$$

and the steering vector is

$$\mathbf{d}(\theta,\omega) = [1, \exp(-j\omega\tau)]^T. \tag{10}$$

The power of array output is then

$$\begin{aligned} P(\theta,\omega) &= \mathbf{d}(\theta,\omega)^H \mathbf{R}(\omega) \mathbf{d}(\theta,\omega) \\ &= r_{11} + r_{22} + 2 \cdot \Re\big\{r_{12} \exp(-j\omega\tau)\big\}, \end{aligned} \tag{11}$$

where $\Re$ indicates the real component.

The first and second terms of (11) do not depend on θ and we only need to consider the third term. The third term of (11), $\hat{P}(\theta,\omega)$, is

$$\begin{aligned} \hat{P}(\theta,\omega) =& P(\theta,\omega) - E[|X_1|^2] - E[|X_2|^2] \\ =& 2\Re\Big[E\big[\sum_{k=1}^{K} H_{1k} H_{2k}^* |S_k|^2 \\ &+ \sum_{l \neq k} H_{1k} H_{2l}^* S_k S_l^*\big] \exp(-j\omega\tau)\Big]. \end{aligned} \tag{12}$$

Therefore, the average of $\hat{P}(\theta,\omega)$ over the frequency bins is

$$\begin{aligned} \hat{P}_{avrg}(\theta) =& \frac{1}{N} \sum_{i=1}^{N} \hat{P}(\theta,\omega_i) \\ =& \frac{1}{N} \sum_{i=1}^{N} \Big[\sum_{k=1}^{K} 2 \cdot E\big[|S_k(\omega_i)|^2\big] \\ & \cdot \Re\big\{H_{1k}(\omega_i) H_{2k}(\omega_i)^* \exp(-j\omega_i\tau)\big\}\Big] \\ &+ \frac{1}{N} \sum_{i=1}^{N} \Big[\sum_{l \neq k} 2 \cdot \Re\big\{H_{1k}(\omega_i) H_{2l}(\omega_i)^* \\ & \cdot E\big[S_k(\omega_i) S_l(\omega_i)^*\big] \exp(-j\omega_i\tau)\big\}\Big], \end{aligned} \tag{13}$$

where N is the number of frequency bins. Since $E[S_k(\omega)S_l(\omega)^*]$ is generally smaller for $k \neq l$ than for $k = l$ we have assumed that the second term in (13) can be set to zero. We can then express (13) as

$$\hat{P}_{avrg}(\theta) \approx \sum_{k=1}^{K} \hat{P}_{avrg}(\theta|\theta_k), \tag{14}$$

where $\hat{P}_{avrg}(\theta|\theta_k)$ is the frequency average of the power of array output from the k'th sound.

As we are only interested in finding the DOAs at this point, we let $H_{1k}(\omega) = \exp(-j\omega\tau_{1k})$ and $H_{2k}(\omega) = \exp(-j\omega\tau_{2k})$. Thus, the frequency average from the k'th sound, $\hat{P}_{avrg}(\theta|\theta_k)$ becomes

$$\begin{aligned} \hat{P}_{avrg}(\theta|\theta_k) =& \frac{2E[|S_k(\omega_i)|^2]}{N} \\ & \cdot \sum_{i=1}^{N} \Re\big\{\exp(-j\omega_i(\tau_{1k} - \tau_{2k})) \\ & \cdot \exp(-j\omega_i\tau)\big\}, \end{aligned} \tag{15}$$

$$\tau_{1k} - \tau_{2k} = \frac{d \sin\theta_k}{c}, \tag{16}$$

$$\tau = \frac{d \sin\theta}{c}, \tag{17}$$

where θ_k is the true direction of the k'th sound position and θ is the steering direction.

2.2. Matching Pursuit to estimate DOA

A matching pursuit algorithm was introduced to decompose any signal into a linear expansion of waveforms [11]. We used a modified matching pursuit algorithm that includes a re-optimization step [12] to decompose the signal into a set of direct and reflected sounds. We define the vector of the angles of i DOAs which are estimated during i iterations as

$$\Theta_i = [\hat{\theta}_1, \cdots, \hat{\theta}_i]^T, \tag{18}$$

where Θ_0 is a vector without any elements. The matching pursuit algorithm for DOA estimation consists of the following steps:

1. Define a dictionary as

$$\mathcal{D} = \{\hat{P}_{avrgn}(\theta|\theta_k)\}_{-\pi/2 < \theta_k < \pi/2}, \tag{19}$$

i.e., an element of family $\mathcal{D}$ is defined as (15) normalized by its norm:

$$\hat{P}_{avrgn}(\theta|\theta_k) = \frac{\hat{P}_{avrg}(\theta|\theta_k)}{\sqrt{\frac{1}{\pi}\int_{-\pi/2}^{\pi/2} |\hat{P}_{avrg}(\theta|\theta_k)|^2 d\theta}}. \tag{20}$$

2. Initialization:

$$e_0(\theta) = \hat{P}_{observed}(\theta) \tag{21}$$

$$i = 1. \tag{22}$$

3. Calculate the residual for all θ_k:

$$e_i(\theta|\theta_k) = e_{i-1}(\theta) - a_{i-1}(\theta_k) \hat{P}_{avrgn}(\theta|\theta_k), \tag{23}$$

where $a_{i-1}(\theta_k)$ denotes the inner product of $e_{i-1}(\theta)$ and $\hat{P}_{avrgn}(\theta|\theta_k)$.

4. Select θ_k (estimate DOA $\hat{\theta}_i$):

$$\hat{\theta}_i = \underset{\theta_k}{\operatorname{argmin}} \sum |e_i(\theta|\theta_k)|^2. \quad (24)$$

5. Re-optimize Θ_i (all DOAs) and calculate the residual $e_i(\theta)$:

$$e_i(\theta) = e_0(\theta) - \sum_{l=1}^{i} \hat{a}(\hat{\theta}_l)\hat{P}_{avrgn}(\theta|\hat{\theta}_l), \quad (25)$$

where $\hat{a}(\hat{\theta}_l)$ is computed from (32).

6. If

$$10 \log \frac{\int e_0^2(\theta)d\theta}{\int e_i^2(\theta)d\theta} < \delta, \quad (26)$$

where δ is the stopping criterion,

$$i = i + 1, \quad (27)$$

go to step 3, or else end the procedure.

2.3. Re-optimization of the DOAs

A high quality, consistent analysis-synthesis method with re-optimization of amplitude and frequency parameters in sinusoidal coding is described by Vos [12]. Here, we use a similar method of re-optimizing the DOAs with a gradient algorithm. We define the vector of the angles of L DOAs as

$$\Theta = [\theta_1, \cdots, \theta_L]^T. \quad (28)$$

The basis vectors and the observed vector are defined as

$$\hat{\mathbf{p}}_{avrgn}(\theta_k) = [\hat{P}_{avrgn}(-\frac{\pi}{2}|\theta_k), \cdots, \hat{P}_{avrgn}(\frac{\pi}{2}|\theta_k)]^T, \quad (29)$$

$$\mathbf{e}_0 = [e_0(-\frac{\pi}{2}), \cdots, e_0(\frac{\pi}{2})]^T, \quad (30)$$

where we have discretized the normalized frequency average of the power of array output of the k'th sound as a function of continuous steering direction variable θ.

For a given set of DOAs, the analysis matrix containing the basis vectors is constructed according to

$$\hat{\mathbf{P}}_{avrgn\Theta} = [\hat{\mathbf{p}}_{avrgn}(\theta_1), \cdots, \hat{\mathbf{p}}_{avrgn}(\theta_L)]. \quad (31)$$

The projection of $\mathbf{e}_0$ onto a space that is defined by the column vectors $\hat{\mathbf{p}}_{avrgn}(\theta_1), \cdots,$ and $\hat{\mathbf{p}}_{avrgn}(\theta_L)$ is

$$\hat{\mathbf{a}} = (\hat{\mathbf{P}}_{avrgn\Theta}^T \cdot \hat{\mathbf{P}}_{avrgn\Theta})^{-1} \cdot \hat{\mathbf{P}}_{avrgn\Theta}^T \cdot \mathbf{e}_0. \quad (32)$$

which is from the least-squares residual. Optimum DOAs are those for which the energy of the projection of $\mathbf{e}_0$ onto the column space of $\hat{\mathbf{P}}_{avrgn\Theta}$ is maximized

$$\underset{\Theta}{\operatorname{argmax}} \{\hat{\mathbf{P}}_{avrgn\Theta} \cdot \hat{\mathbf{a}}\}^T \cdot \{\hat{\mathbf{P}}_{avrgn\Theta} \cdot \hat{\mathbf{a}}\}$$
$$= \underset{\Theta}{\operatorname{argmax}}\, \mathbf{e}_0^T \mathbf{P}_\Theta \mathbf{e}_0, \quad (33)$$

where we define the projection matrix as

$$\mathbf{P}_\Theta = \hat{\mathbf{P}}_{avrgn\Theta} \cdot (\hat{\mathbf{P}}_{avrgn\Theta}^T \cdot \hat{\mathbf{P}}_{avrgn\Theta})^{-1} \cdot \hat{\mathbf{P}}_{avrgn\Theta}^T. \quad (34)$$

We used a gradient search based on Newton's method to find the local maximum of $\mathbf{e}_0^T \mathbf{P}_\Theta \mathbf{e}_0$ in the neighborhood of a set of initial DOAs. At iteration m, Newton's method defines the updated DOAs as

$$\Theta^{(m)} = \Theta^{(m-1)} + \mathbf{H}_{\Theta^{(m-1)}}^{-1} \mathbf{g}_{\Theta^{(m-1)}}, \quad (35)$$

where $\mathbf{g}_\Theta$ and $\mathbf{H}_\Theta$ represent the gradient vector and the Hessian matrix of the energy of the projection:

$$\mathbf{g}_\Theta = \frac{\partial}{\partial \Theta} \mathbf{e}_0^T \mathbf{P}_\Theta \mathbf{e}_0, \quad (36)$$

$$\mathbf{H}_\Theta = \frac{\partial^2}{\partial \Theta \partial \Theta^T} \mathbf{e}_0^T \mathbf{P}_\Theta \mathbf{e}_0. \quad (37)$$

3. IMPULSE RESPONSE ESTIMATION AND UNMIXING

If we assume that the estimated L DOAs have come from two different sources, we can classify L different sounds into two categories based on their cross correlation. If L_1 sources are classified into category 1, and L_2 sources are classified into category 2, the mixing matrix can be estimated in the time domain as

$$\hat{\mathbf{A}} = \begin{bmatrix} \hat{A}_{11}(t) & \hat{A}_{12}(t) \\ \hat{A}_{21}(t) & \hat{A}_{22}(t) \end{bmatrix}, \quad (38)$$

$$\hat{A}_{11}(t) = \sum_{i=1}^{L_1} a_{1i} \cdot \delta\big(t - (D_0 + \frac{\tau_{1i}}{2} + D_{1i})\big), \quad (39)$$

$$\hat{A}_{12}(t) = \sum_{i=1}^{L_2} a_{2i} \cdot \delta\big(t - (D_0 + \frac{\tau_{2i}}{2} + D_{2i})\big), \quad (40)$$

$$\hat{A}_{21}(t) = \sum_{i=1}^{L_1} a_{1i} \cdot \delta\big(t - (D_0 - \frac{\tau_{1i}}{2} + D_{1i})\big), \quad (41)$$

$$\hat{A}_{22}(t) = \sum_{i=1}^{L_2} a_{2i} \cdot \delta\big(t - (D_0 - \frac{\tau_{2i}}{2} + D_{2i})\big), \quad (42)$$

where τ_{1i} and τ_{2i} are the time delays between microphones for each DOA i, D_0 is an initial delay (which is arbitrary), D_{1i} and D_{2i} are the time lags from the first sounds (which are obtained from calculating cross correlations of sounds and the first sound), and a_{1i} and a_{2i} are the amplitudes of sounds obtained from DOA estimation results ($\hat{a}(\hat{\theta}_l)$ in (25)).

We computed unmixing matrix $\mathbf{B}$ as the inverse of the mixing matrix

$$\mathbf{B} = \hat{\mathbf{A}}^{-1}. \quad (43)$$

4. EXPERIMENTS

We separated mixture sources both for artificial and real-world data. The distance between microphones was 5 cm under both conditions. The artificial conditions are in Fig. 1. The speech signals were assumed to arrive from four directions. This is equivalent to a mixture with one direct and one reflected sound for each independent source. The real-world data were recorded in our office, which had a reverberation time of 0.76 s. Sources were located at -30 degrees and 45 degrees. We used a data sampling frequency of 44.1 kHz, a frame length of 46 ms, and a frame update of 23 ms.

We compared our results with those obtained from the ICA method of Kurita [5]. We set the number of iterations to 1000 and the step-size parameter to 0.001 for this conventional ICA based BSS. The stopping criterion was set to 30 dB for the new method. To evaluate performance, we used noise reduction rate (NRR), which is defined as the output signal to noise ratio (SNR) in dB minus the input SNR in dB [9].

Figure 2 has the matching pursuit iteration for the artificial data. The x-axis is the direction of arrival, and the y-axis is the power of the arriving sound. The top curve is the observed power e_0. The second curve is the residual e_1 after the first DOA is estimated. The third, fourth, and last curve are the residual e_2, e_3, and e_4. Tables 1 and 2 list the experimental results. We found ten sounds in the real-world data for the stopping criterion of 30 dB. The residual e_{10} was almost flat. Separation with proposed method was superior to that of conventional ICA based BSS both for artificial and real-world data experiments. Separation improvement

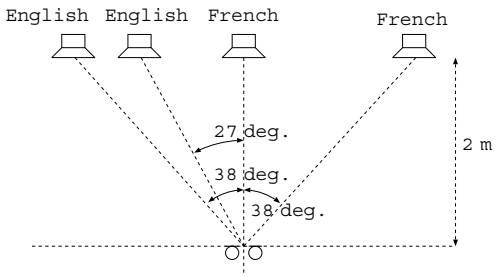

Fig. 1. Mixture conditions.

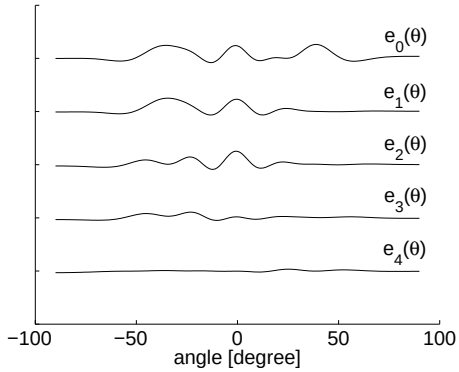

Fig. 2. Matching pursuit iterations.

Table 1. NRR values for artificial data

Method	NRR [dB]
Conventional ICA	18.7
Proposed method	25.5

Table 2. NRR values for real-world data

Method	NRR [dB]
Conventional ICA	3.98
Proposed method	4.34

for the real-world data experiment is less than that for the artificial experiment, since estimation of reflected sounds was insufficient. The additional improvement could be achieved by narrowing the beam in DOA estimation.

5. CONCLUSION

In this paper, we proposed a new BSS method that uses spatial information derived from the results of DOA estimates for each direct and reflected sound obtained by means of beam forming. Its main advantage is that we can estimate the mixing system for direct and early reflected sounds and separate sounds through a suitable technique with sound source separation in the real world. A matching pursuit algorithm that includes a re-optimization step for each iteration was used for DOA estimates and could estimate these correctly.

The source separation with the method we proposed was better than that obtained by ICA both in artificial and real-world data experiments. It improved the noise reduction rate by about 7 dB in an artificial data experiment and by about 0.4 dB in a real-world data experiment. We expect that additional improvement can be achieved in real-world cases if the accuracy of estimation of reflected sounds is increased. This could for instance be achieved by narrowing the beam in DOA estimation. We are currently working on this.

6. ACKNOWLEDGEMENT

The authors express their gratitude to Prof. Bastiaan Kleijn and Prof. Arne Leijon of KTH for many fruitful discussions and suggestions.

7. REFERENCES

[1] J.-F. Cardoso, "Blind signal separation: Statistical principles," *Proc. IEEE*, vol. 86, no. 10, pp. 2009–2025, Oct. 1998.

[2] S. Amari and A. Cichocki, "Adaptive blind signal processing - neural network applications," *Proc. IEEE*, vol. 86, no. 10, pp. 2026–2048, Oct. 1998.

[3] Te-Won Lee, *Independent component analysis. Theory and applications*, Kluwer Academic Publishers, Boston, 1998.

[4] S. Ikeda and N. Murata, "A method of ICA in time-frequency domain," in *Proc. Int. Workshop Independent Comp. Analysis Blind Sign. Separation (ICA'99)*, 1999, pp. 365–371.

[5] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda, and F. Itakura, "Evaluation of blind signal separation method using directivity pattern under reverberant conditions," in *Proc. IEEE Int. Conf. Acoust., Speech, Sign. Process.*, Istanbul, 2000, vol. 5, pp. 3140–3143.

[6] H. Saruwatari, T. Kawamura, and K. Shikano, "Blind source separation for speech based on fast-convergence algorithm with ICA and beamforming," in *Proc. Eurospeech*, Aalborg, 2001, vol. 4, pp. 2603–2606.

[7] A. Jourjine, S. Rickard, and Ö. Yilmaz, "Blind separation of disjoint orthogonal signals: demixing n sources from 2 mixtures," in *Proc. IEEE Int. Conf. Acoust., Speech, Sign. Process.*, Istanbul, 2000, vol. 5, pp. 2985–2988.

[8] R. Gribonval, "Sparse decomposition of stereo signals with matching pursuit and application to blind separation of more than two sources from a stereo mixture," in *Proc. IEEE Int. Conf. Acoust., Speech, Sign. Process.*, Orlando, 2002, vol. 3, pp. 3057–3060.

[9] S. Araki, S. Makino, T. Nishikawa, and H. Saruwatari, "Fundamental limitation of frequency domain blind source separation for convolutive mixture of speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Sign. Process.*, Salt Lake City, 2001, vol. 5, pp. 2737–2740.

[10] D. H. Johnson and D. E. Dudgeon, *Array signal processing: Concepts and techniques*, Prentice-Hall, Inc., New Jersey, 1993.

[11] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.

[12] K. Vos, R. Vafin, R. Heusdens, and W. B. Kleijn, "High-quality consistent analysis-synthesis in sinusoidal coding," in *Proc. 1999 Audio Eng. Soc. 17th Conf. "High Quality Audio Coding"*, 1999, pp. 244–250.